Do Llamas Understand the Periodic Table?

Ge Lei
Dyson School of Design Engineering
Imperial College London
SW7 2AZ, United Kingdom
g.lei23@imperial.ac.uk

Samuel J. Cooper
Dyson School of Design Engineering
Imperial College London
SW7 2AZ, United Kingdom
samuel.cooper@imperial.ac.uk

Abstract

Large Language Models (LLMs) show impressive capacity to synthesize scientific knowledge but struggle with basic arithmetic, raising concerns about reliability. As materials science increasingly leverages LLMs for hypothesis generation, it is essential to understand how they encode specialized knowledge. Here, we investigate how the open-source Llama series of LLMs represent the periodic table of elements. We observe a 3D spiral structure in the hidden states of LLMs that aligns with the conceptual structure of the periodic table, suggesting that LLMs can reflect the geometric organization of scientific concepts learned from text. Linear probing reveals that middle layers encode continuous, overlapping attributes that enable indirect recall, while deeper layers sharpen categorical distinctions and incorporate linguistic context. These findings suggest that LLMs represent symbolic knowledge not as isolated facts, but as structured geometric manifolds that intertwine semantic information across layers. We hope this inspires further exploration into the interpretability mechanisms of LLMs within chemistry and materials science, enhancing trust of model reliability, guiding model optimization and tool design, and promoting mutual innovation between science and AI.

1 Introduction

Large Language Models (LLMs) have demonstrated a notable capacity to synthesize and generate insights from vast amounts of expert knowledge, drawing attention across multiple scientific domains [Wysocki et al., 2024, Lei et al., 2024]. Yet, despite their impressive capabilities, researchers have observed their surprising inability to reliably perform seemingly straightforward tasks, such as basic arithmetic operations [Qian et al., 2022, Baeumel et al., 2025, Gambardella et al., 2024]. This phenomenon highlights an important aspect of LLMs: their fundamental reliance on learned patterns and probabilistic predictions based on token embeddings, rather than explicit arithmetic operations. Consequently, simple numerical tasks, effortlessly handled by even the most rudimentary calculators with orders of magnitude less computation, remain challenging and error-prone for these sophisticated LLMs.

In parallel, interest is rapidly growing in leveraging LLMs within the materials sciences community. Recent research has proposed intriguing applications such as laboratory orchestration [Sim et al., 2024, Darvish et al., 2025, Olowe and Chitnis, 2025], hypothesis generation [Liu et al., 2025a, Kumbhar et al., 2025, Bazgir et al., 2025a,b], and complex materials property prediction [Liu et al., 2025b, Rubungo et al., 2025]. However, despite convincing demonstrations reported in numerous studies, there remains skepticism regarding the reliability and trustworthiness of these systems in rigorous scientific research. One particular risk is that large language models are designed during training to generate responses in ways that align with the user's expectations. This can make the answers appear more authoritative than they actually are, potentially giving users a false sense of confidence in the output, even when the underlying information may be incorrect or fabricated [Steyvers et al., 2025].

Accepted at the AI for Science Workshop (NeurIPS 2025)

This raises a critical question: Can LLMs be trusted to accurately represent chemical information and serve as scientific foundation models? To approach these questions, understanding internal representations is critical. If LLMs simply memorize isolated facts, they would need constant task-specific fine-tuning, effectively functioning as many small models. By contrast, if they organize complex knowledge into structured forms, this would suggest an ability to extract and generalize abstract regularities. More ambitiously, if these representations are geometry-aware and aligned with physical laws, this may suggest that LLMs can capture aspects of universal scientific regularities. Such representations would support compression and generalization, enabling inference of unmentioned properties and transfer across tasks, thereby increasing their potential to contribute meaningfully to scientific research.

In this work, we investigate whether the prominent open-source LLMs, Llamas [Grattafiori et al., 2024], store chemical knowledge in a structured and rational manner, whether fragmented into isolated clusters of disconnected facts, or interconnected through rational webs of structured knowledge. We delve into how LLMs encode and recall such knowledge through layer-wise, geometry-aware representations. The contributions of our study are:

1. We report the first observation of a 3D spiral structure in LLM hidden states that organizes chemical elements in alignment with the structure of the periodic table (Sec 3).

2. To our knowledge, this is the first to compare regression and classification probing, showing that middle layers encode continuous attribute structure, while later layers sharpen boundaries for fine-grained decisions (Sec 4.1).

3. We show that linguistic structure increasingly shapes knowledge representations in later layers (Sec 4.2).

4. We find that LLMs recall related attributes through strong linear associations in middle layers, which weaken in deeper layers (Sec 5).

2 Preliminaries

Our study only focuses on how reliably acquired knowledge (*i.e.* things we're confident the model knows) is represented within LLMs, and excludes hallucinations or information not in the training set. We use the properties of chemical elements in the periodic table as a case study due to their frequent occurrence in training data, well-defined attributes, quantifiable properties, and making them an ideal subject for this investigation. We adopt Llama series models [Touvron et al., 2023, Grattafiori et al., 2024] in this study. To study how LLMs represent attributes across layers, we construct a prompt dataset based on a set of attributes ($A = \{A_j\}_{j=1}^{M}$, such as 'atomic number' or 'group') and a set of elements ($X = \{X_i\}_{i=1}^{N}$, constituting the first 50 elements, such as 'Mg' or 'Al'). For linguistic diversity, we incorporate predefined template sets: $T^{\text{cont}} = \{T_k^{\text{cont}}\}_{k=1}^{11}$ for continuation-style prompts and $T^{\text{ques}} = \{T_k^{\text{ques}}\}_{k=1}^{11}$ for question-style prompts, with 11 templates in each. In the continuation-style templates, the next output token would be the factual knowledge directly such as:

$$T_2^{\text{cont}}(A_j, X_i) = \text{`}X_i\text{'s } A_j \text{ is'}$$

In question-style templates, the next output token is typically a syntactic word like 'The', which ensures the grammatical structure is correct, such as:

$$T_1^{\text{ques}}(A_j, X_i) = \text{`What is the } A_j \text{ of } X_i\text{?'}$$

$$T_2^{\text{ques}}(A_j, X_i) = \text{`Which value represents } X_i\text{'s } A_j\text{?'}$$

By substituting each element and attribute (X_i, A_j) into these templates, we generate prompts:

$$p_{i,j,k} = T_k(X_i, A_j)$$

Each prompt $p_{i,j,k}$ can then be fed into LLMs to study the corresponding residual streams at different layers. Last-token residual streams capture the full prompt context in decoder-only models with masked attention, as they integrate information from all preceding tokens. For each layer l, we collect last-token residual streams $\mathbf{h}_{i,j,k}^{(l)}$ from prompts $p_{i,j,k}$ across all elements and templates (see Appendix A for details).

3 Geometric relationships among attributes

In materials science, a spiral trend emerges from the periodic variation in valence electron configurations as atomic number increases. By arranging elements sequentially and mapping their properties in a polar coordinate system, this periodicity becomes visually apparent as a spiral. We investigate whether LLMs (that have all been exposed to extensive data on the properties of elements during training), inherently capture these physical periodicities and reflect similar spiral structures in their learned embeddings. We hypothesize that attributes in LLMs exist in a high-dimensional space, manifesting as linear, circular, or spiral patterns based on their structure, and then proceed to validate these geometries.

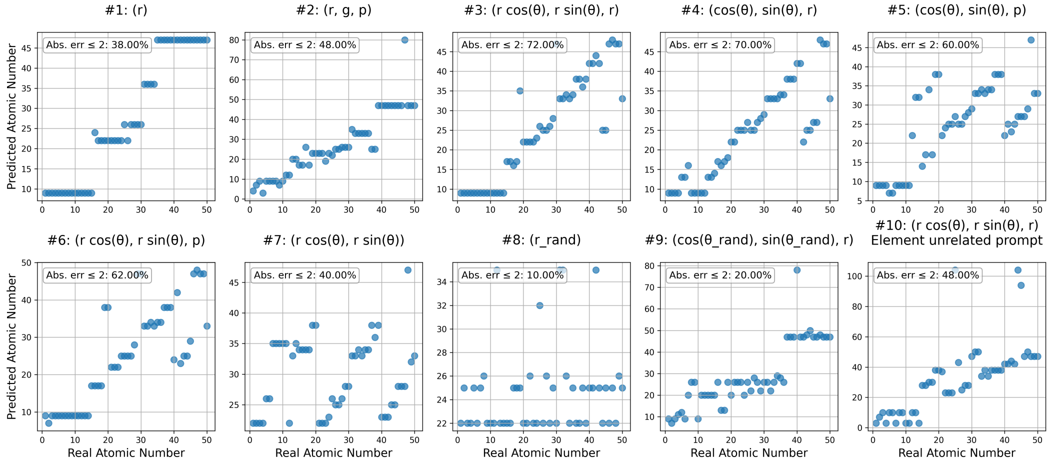


Figure 1: Residual Stream Patching Results for Layer 20 in Meta-Llama-3.1-70B. The model's predictions are evaluated after replacing the residual stream of the 'element' token at the last token position with the predicted residual stream $\hat{\mathbf{h}}_0^{\text{pred},(20)}$.

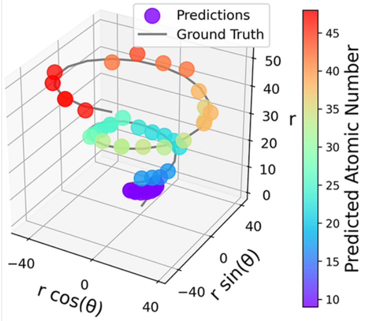


Figure 2: Predicted atomic numbers after intervention in 3D spiral space $(r\cos\theta, r\sin\theta, r)$. Colored points indicate the tokens with highest logits.

Inspired by Engels et al. [2024], we map the last-token residual streams $\mathbf{h}^{(l)} \in \mathbb{R}^k$ at layer l to a geometric space $f(r, g, p)$, which encodes atomic number r, group g, and period p. To learn this mapping, we first reduce the dimensionality of the residual streams to 30 using PCA, denoted as $\mathbf{P}(\mathbf{h}^{(l)})$, and fit a linear projection using all 50 elements except one held-out target:

$$\mathbf{W}^{(l)}, \mathbf{b}^{(l)} = \arg\min_{\mathbf{W}', \mathbf{b}'} \sum_{i \neq 0} \left\| \mathbf{W}' \mathbf{P}(\mathbf{h}_i^{(l)}) + \mathbf{b}' - f_i \right\|_2^2$$

where $\mathbf{W}^{(l)} \in \mathbb{R}^{d' \times 30}$, $\mathbf{b}^{(l)} \in \mathbb{R}^{d'}$, and $f_i = f(r_i, g_i, p_i)$ denotes the mapping of the i-th element in the geometric space.

To perform the intervention, we compute the centroid of the PCA-reduced residual streams for the remaining $N = K - 1$ elements:

$$\bar{\mathbf{h}}^{(l)} = \frac{1}{N} \sum_{i \neq 0} \mathbf{P}(\mathbf{h}_i^{(l)})$$

then map it to the geometric space: $\mathbf{z} = \mathbf{W}^{(l)}\bar{\mathbf{h}}^{(l)} + \mathbf{b}^{(l)}$. Let $f_0 = f(r_0, g_0, p_0)$ denote the target element's embedding in the geometric space. The deviation $f_0 - \mathbf{z}$ is projected back to the residual stream space using the pseudo-inverse of $\mathbf{W}^{(l)}$, giving the predicted (intervened) residual stream:

$$\hat{\mathbf{h}}_0^{\text{pred},(l)} = \mathbf{P}^{-1}\left(\bar{\mathbf{h}}^{(l)} + \left(\mathbf{W}^{(l)}\right)^+ (f_0 - \mathbf{z})\right)$$

Importantly, the model never accesses the original residual stream of the target element; the predicted residual stream is computed solely from its geometric representation and the residual streams of other elements. During inference, we replace the residual stream of 'element' (last token position) in the 20th layer[1] with $\hat{\mathbf{h}}_0^{\text{pred},(20)}$, using the prompt *'In the periodic table, the atomic number of element'*. We then evaluate whether the model can correctly output the target token without ever seeing its original residual stream.

We evaluate the effectiveness of different geometric spaces for interventions, including linear, 2D spiral, and 3D spiral (i.e. conical helix) geometries, as shown in Figure 2 . Angular variables $\theta = \frac{2\pi g}{18}$ are used to capture periodic relationships. To test the impact of disrupted geometry, two random spaces are introduced: in Space 8, atomic numbers r are shuffled; in Space 9, θ is randomly permuted. Additionally, in Space 10, the prompt *'In numbers, the Arabic numeral for number'* generates numbers 1–50, testing whether periodic patterns emerge without explicit element references. We designed this control to examine whether the observed geometric shapes arise from element-related knowledge or simply from numerical sequences.

Effective residual stream patching suggests that the target space $f(r, g, p)$: 1) retains sufficient information for accurate reconstruction during transformations with the residual stream space, and 2) preserves geometric structures similar to those in the residual stream space to ensure valid adjustments in the high-dimensional space.

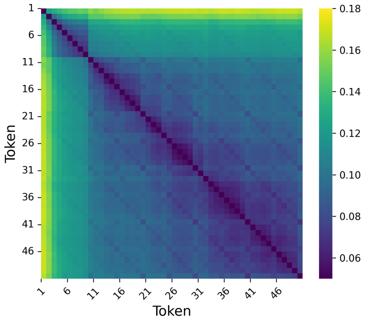


Figure 3: Euclidean distance heatmap of approximated vector representations for numeric tokens (1-50) in the hidden space of the last layer in Meta-Llama-3.1-70B.

Patching results for Meta-Llama-3.1-70B are shown in Fig. 1, with detailed values in Table B.1 (Appendix). Results show that intervention can be applied in various geometric spaces, with some performing significantly better. Spaces such as $(\cos\theta, \sin\theta, r)$ and $(r\cos\theta, r\sin\theta, r)$ over 70% predictions of the atomic number have an absolute error within 2, suggesting the potential existence of latent 3D structures in LLMs resembling 3D spirals. Fig. 2 illustrates the LLM's output post-intervention in 3D spiral geometry. Additional geometric analyses are in Appendix B.2. Randomly generated prompts perform very poorly, which is expected given their lack of coherent semantic structure and context, but even element unrelated prompts with clear linguistic form also yield poor performance. This suggests that the geometry of the embedding space is not merely tied to numerical correlations or surface-level semantics, but is inherently aligned with the background knowledge invoked by the prompt, reflecting real-world knowledge structures.

In a concurrent study, Kantamneni and Tegmark [2025] observed spiral-like structures in number space with periods of 2, 5, 10, and 100, likely reflecting common human conventions in numerical representation. In contrast, our model exhibits a distinct 18-period spiral [2] aligned with the periodic structure of chemical elements. This representation performs notably worse for ordinary numbers without elemental context (which aligns with their observation that the 18-period does not prominently emerge), indicating that such geometric patterns emerge from underlying physical or semantic regularities rather than arbitrary structures.

[1] See Appendix B.1 for details on intervention performance. Interventions become effective from layer 20 onward.

[2] In our study, each period is arranged on an 18-sector scale corresponding to the 18 groups; in the first three periods, only 2, 8, 8 sectors are filled, with the remaining sectors left empty. Thus, the finer 2/8/8/18 Period is preserved within the 18-sector framework.

In the intervention experiments, it is actually not obvious whether a smaller numerical difference between the output token and the true value always implies smaller error. To investigate this, we project token IDs for numbers 1–50 into the last hidden layer using the pseudoinverse of the vocabulary projection matrix $\mathbf{W}_{\text{vocab}}^{+}$. This operation reconstructs an approximation of the hidden representations that would produce these token IDs as logits. Fig. 3 shows that smaller numerical differences generally correspond to closer representations, while larger differences often result in inconsistent distances, reflecting the model's difficulty with numerical consistency over larger gaps. For instance, the vector for '1' is closer to '2' than to '5', while the distances between '10' and '40' is closer than between '10' and '21'. In the intervention, when the predicted value is close to the true value, hidden logits align well with true logits, suggesting higher accuracy. However, large numerical deviations cannot fully capture prediction errors, so we evaluate results using an absolute error threshold (≤ 2) in Fig. 1, representing a small distance.

4 Direct attribute recall

In the previous section, we observed that elemental knowledge in LLMs forms a 3D spiral structure. Interestingly, although prompts mentioned only atomic numbers, the embeddings also reflected elemental groups, suggesting that LLMs retrieve both explicitly requested and implicitly related attributes. To better understand these mechanisms, this section investigates direct attribute knowledge recall and Sec. 5 will explores how LLMs access related but unprompted knowledge.

4.1 From continuity to boundary sharpening

Some elemental attributes, such as group and period, naturally exist in both categorical and numerical forms. This duality enables both classification and regression probing, allowing for direct comparisons that have been underexplored in prior work, which often focused exclusively on a single type.

To examine how LLMs access explicitly mentioned knowledge, we use the last-token residual stream from the continuation style prompt $\mathbf{h}_j^{(l)} \in \mathbb{R}^k$ as the representation of attribute A_j, and fit a linear probe to predict its corresponding values via:

$$f_j^{(l)}(\mathbf{h}_j) = \mathbf{W}_j^{(l)}\mathbf{h}_j^{(l)} + \mathbf{b}_j^{(l)}$$

For categorical attribute forms (*e.g.*, Category, Group, Period), $\mathbf{W}_j^{(l)} \in \mathbb{R}^{|C_j|\times k}$, $\mathbf{b}_j^{(l)} \in \mathbb{R}^{|C_j|}$. Predictions are made by:

$$\hat{y}^{(l)} = \arg\max_{c\in C_j}[f_j^{(l)}(\mathbf{h}_j^{(l)})]_c.$$

For continuous attributes, we perform scalar regression by setting $\mathbf{W}_j^{(l)} = \mathbf{w}_j^{(l)\top}$, $\mathbf{w}_j^{(l)} \in \mathbb{R}^k$, $\mathbf{b}_j^{(l)} \in \mathbb{R}$, yielding:

$$\hat{y}^{(l)} = \mathbf{w}_j^{(l)\top}\mathbf{h}_j^{(l)} + b_j^{(l)}$$

Probes are trained using 5-fold cross-validation on last-token residual streams. We use a linear Support Vector Machine (SVM) for categorical tasks and Support Vector Regression (SVR) with a linear kernel for continuous tasks. The resulting classification accuracies and regression R^2 scores are shown in Fig. 4, with best-layer results provided in appendix F.5.

Regression probes reveal that continuous numerical features are effectively represented in intermediate layers, as indicated by high R^2 values (while not reaching 1, see Appendix F.1). These intermediate layers sometimes even outperform the final layers, suggesting that numerical knowledge is already encoded before the final output stage. This aligns with findings by [Meng et al., 2022], which show that factual knowledge recall is already mediated by intermediate MLP layers.

In classification probes, intermediate layers perform similarly or even better than final layers for clearly distinct non-numerical categories (*e.g.*, metal vs. non-metal), aligning with prior work [Nanda et al., 2023]. However, they significantly underperform in fine-grained numerical classification, *e.g.*, Period accuracy drops from ~ 1.0 (final) to ~ 0.7, and Group from ~ 1.0 to ~ 0.6.

This suggests that while intermediate layers already encode meaningful numerical structure, additional processing in later layers is required to sharpen boundaries and support accurate fine-grained

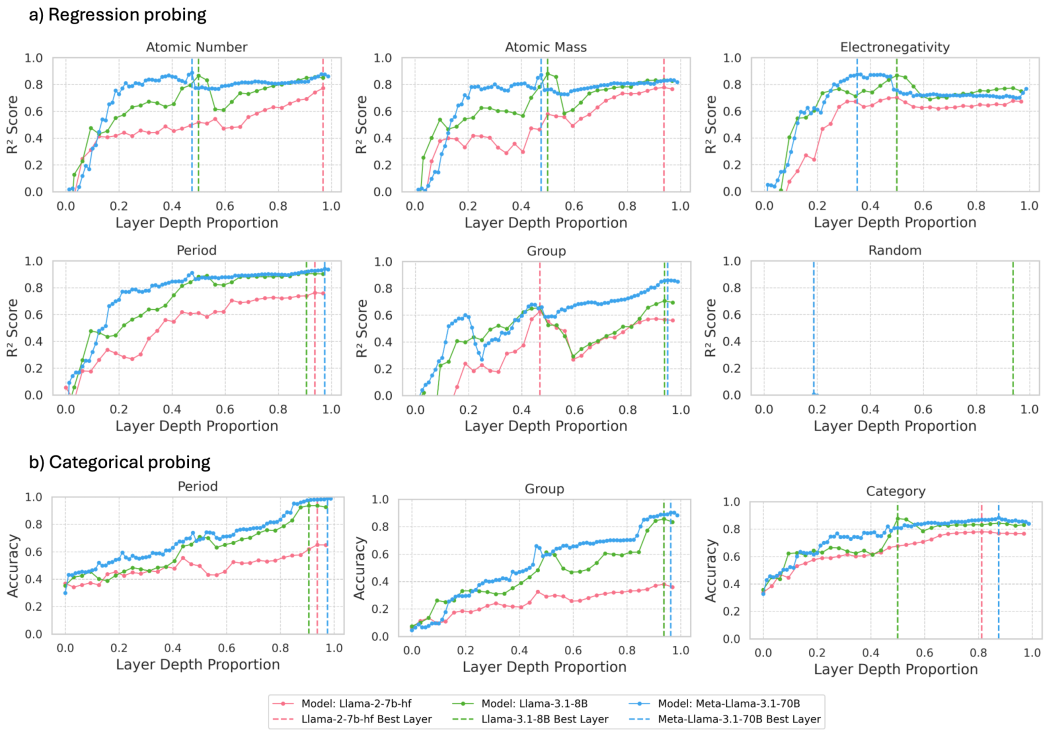


Figure 4: Linear probing results on last token across layers. (a) Regression (R^2) for numerical attributes and a random baseline. (b) Classification (accuracy) for categorical attributes. All results use 5-fold cross-validation on last-token residual streams.

classification. This aligns with intuition: later layers prepare for discrete token outputs, where clearer classification boundaries must emerge. As shown in Appendix Fig. F.1, the confusion matrix from Layer 40 (70B middle layer) is not perfectly accurate, but most misclassifications fall near the diagonal, further demonstrating that intermediate layers encode coherent numerical structure, albeit with blurred categorical boundaries. These observations may provide useful insights for choosing between intermediate and later-layer embeddings in downstream tasks.

Notably, Llama2 7B shows low accuracy (<0.4) on Group classification compared to Llama3.1 8B (>0.8) (but similar performance in Group regression probing) potentially due to its single number tokenization (splitting numbers like ‘12’ into ‘1’ and ‘2’), which may cause confusion between the representations of output tokens like ‘12’ and ‘1’. In contrast, Llama 3 uses separate tokens for numbers below 1000.

4.2 Higher Language Sensitivity in Later Layers

The sharpening of numerical representations into categorical boundaries in later layers suggests that these layers might be shaped by the expected output tokens. This raises a question: does the linguistic structure influence the factual representations across layers?

We compared question-style and continuation-style prompts using linear regression probes. Continuation prompts generally lead to direct generation of fact-related tokens, whereas question-style prompts tend to introduce syntactic fillers (*e.g.*, ‘The’) and are more influenced by superficial language patterns.

Fig.5 reports the average delta R^2 across five attributes, with per-attribute results shown in Fig. F.2 (Appendix). As analyses in earlier sections show stronger semantic signals and higher R^2 in mid-to-late layers, we focus on depths 0.6–1.0. ΔR^2 increases in the mid-to-late layers, indicating a growing gap between prompt types. Among the 15 attribute–model combinations (3 models × 5 attributes), 12 show a significant increasing trend (FDR-corrected $p < 0.05$), with a median Mann–Kendall τ of 0.55 (Appendix F.3).

The results indicate that, as depth increases, question prompts become progressively less effective than continuation prompts at encoding factual attributes, hinting that the prompt's linguistic structure exerts a stronger influence on representations in deeper layers. Interestingly, the larger models show a slower increase in ΔR^2 across layers than the smaller models, suggesting they maintain more stable factual representations across prompt types and thus exhibit a smaller distinction between continuation and question prompts.

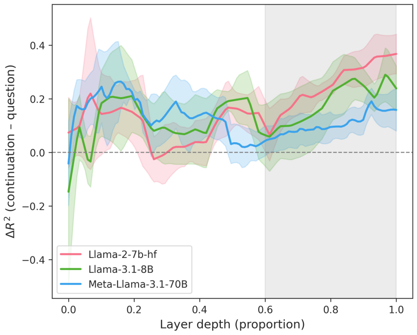


Figure 5: Average ΔR^2 across five attributes, with 95% confidence interval shaded. $\Delta R^2 = R^2_{\text{cont}} - R^2_{\text{ques}}$.

The rising ΔR^2 suggests that deeper layers increasingly blend factual content with linguistic structure to prepare the final tokens. To further test this, we applied the logit lens [nostalgebraist, 2020] and tuned-lens [Belrose et al., 2023]. These analyses estimate the token distribution each layer would produce if decoding were halted at that depth, and show that the correct numerical token becomes highly ranked only in the later layers (Appendix D). Complementary attention statistics (Appendix C) reveal that mid-layers focus tightly on the factual token, whereas later layers spread attention over a wider context patterns consistent with increased syntactic and contextual integration.

5 Indirect attribute recall

In the previous section, we analyzed direct recall of explicitly mentioned attributes across layers. Our earlier geometric analysis showed that LLMs can also recall related attributes that are not explicitly mentioned. In this section, we explore how related but unmentioned attributes are recalled.

5.1 Middle Layers excel at indirect recall

We conducted experiments using linear probing to examine the relationships between distinct attributes. Specifically, we extracted last-token residual streams from continuation prompts that mention attribute A_{j_1} (matching) or a different attribute A_{j_2} (non-matching), i.e. seeing if we can extract information that was not explicitly requested in the prompt. We also extracted the residual stream at the element token position, before any attribute is introduced (no mention). Separate probes were trained for each residual stream dataset, always using labels of attribute A_{j_1} as targets. To avoid confounding factors, we selected six attribute pairs without direct linear relationships for non matching probe (see Appendix F.4.1). Average R^2 curves for all attributes are shown in Fig. 6; detailed case-wise linear probing results appear in Appendix Figs. F.3 and F.4.

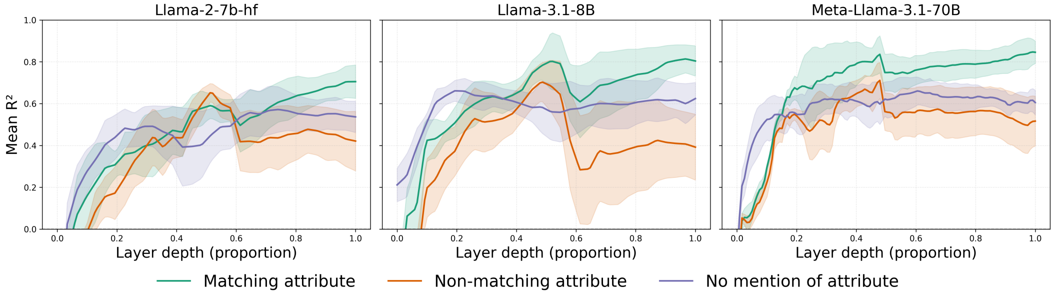


Figure 6: Average R^2 scores from regression probing across layers for three prompt types: matching, non-matching, and no-mention. All probes predict a fixed target attribute A_{j_1}; no-mention uses element token residual streams before any attribute appears. Shaded areas show 95% confidence intervals.

Attribute information was detectable across all prompt styles. Intuitively, matching prompts should perform best by providing explicit cues, no-mention comes next as it relies on inference, and non-matching prompts perform worst due to misleading signals. Surprisingly, at intermediate layers (around 0.5 depth), non-matching prompts yielded higher linear R^2 scores than no-mention prompts, suggesting stronger inter-attribute interactions at these depths. This may reflect entangled representations between related attributes, which we analyze further in Sec. 5.2.

Beyond 60% depth, performance follows the expected trend: matching > no-mention > non-matching. The gap between matching and non-matching prompts increases steadily from 0.6 to 1.0 depth. Across 15 model–attribute tests, 14 exhibited statistically significant divergence (FDR corrected $p < 0.05$), with a median Mann–Kendall τ of 0.77 (Appendix Fig.F.5, Table F.3). It suggests that attribute representations become more specialized and context-sensitive in deeper layers. Further analyses in Sec. 5.2 provide a more direct explanation, examining how structural relationships between attributes contribute to this layered specialization.

The fact that the 'no-mention' prompts perform best in the early layers may seem counterintuitive; however, this is likely because, unlike the other two scenarios, in the 'no-mention' case, the last token is the element itself, which may aid recall. In contrast, matching prompts extract residual streams at the final token (such as 'is'), requiring holistic semantic understanding. As layer depth increases, semantic clarity improves, enhancing explicitly mentioned attributes and reversing this initial trend.

5.2 Stronger linear correlations in middle layers

To explicitly capture relationships between attribute representation, we train a linear mapping from the representation of attribute A_{j1} to attribute A_{j2} at each model layer. Specifically, we utilize the final residual streams from a fixed prompt template (after applying PCA to reduce the dimensionality to 20). The mapping performance is evaluated using R^2 scores obtained via 5-fold cross-validation.

Fig. 7 illustrates the variation of R^2 scores across layers for different attribute pairs. In early layers, R^2 scores are high; however, this observation alone does not necessarily indicate meaningful attribute-level relationships, as initial representations are predominantly sensitive to token-level similarity. Due to the use of a fixed input template, the resulting inputs exhibit substantial token-level overlap.

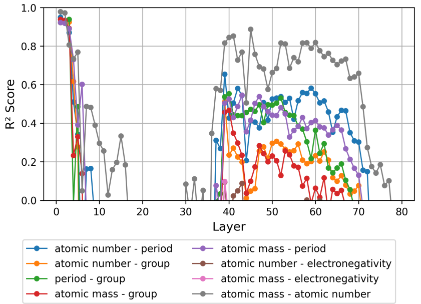


Figure 7: R_2 scores across layers in Meta-Llama-3.1-70B for linear mappings between attribute pairs using the final residual stream from a fixed prompt.

In the intermediate layers, where concept-level understanding is evident (as shown by t-SNE and linear probing), we observe a peak in R^2 scores. This indicates that even simple linear models can effectively capture relationships between different attributes, reflecting their connection in the learned representation space. This also explains why prompts with non-matching attributes outperform those with no attribute mention at these layers in the last section Sec 5.1. In deeper layers, R^2 scores decline, suggesting a shift toward specialized representations. Similar conclusions from the linear probing weight analysis further support this, as shown in Appendix E.1.

6 Discussion and conclusions

This study highlights that despite their exclusive reliance on textual training data, LLMs internally develop structured representations aligning closely with scientific knowledge. Specifically, we observe a 3D spiral structure within the hidden states of LLMs that mirrors the conceptual organization of the periodic table, indicating the models' implicit grasp of domain-specific regularities without explicit supervision.

Probing experiments reveal that the encoding of chemical knowledge evolves across model depth: middle layers encode continuous, overlapping attribute subspaces suitable for coarse categorization, while deeper layers sharpen decision boundaries and integrate linguistic structure. In addition, we find that related attributes are strongly linearly associated in middle layers, enabling indirect recall.

Our results demonstrate that symbolic scientific knowledge, particularly in chemistry, is represented within LLMs as coherent, geometry-aware manifolds where conceptual information is systematically intertwined across model layers. Furthermore, this geometric structure aligns with the laws observed in the physical world, indicating that knowledge within LLMs is not arbitrary, but rather organized and reflective of inherent natural order. Moreover, it is unsurprising that these large models discover meaningful relationships between concepts and these must often represent efficient compression.

We hope this work inspires further investigation into how LLMs represent and reason about scientific knowledge, such as materials property prediction, and informs the design of downstream embedding-based tasks. We believe interpretability in LLMs is essential for AI safety, reducing unintended behaviors and building trust. Understanding how knowledge is stored and recalled across layers can inspire more interpretable, efficient models, advance knowledge editing and scientific discovery.

Limitations. Our prompts have specifically targeted chemical elements in the periodic table; future studies could expand this to include other chemical structures and properties. The hypothesis-driven validation of geometric structures may oversimplify LLMs' non-linear interactions.

Acknowledgements

SJC was supported by the FULL-MAP project, which is from the European Union's research and innovation programme Horizon Europe under the grant agreement No. 101192848 and LG was supported by the Imperial Lee Family Scholarship. We would like to thank the members of the TLDR group for their valuable comments and insightful discussions.

References

Tanja Baeumel, Josef van Genabith, and Simon Ostermann. The lookahead limitation: Why multi-operand addition is hard for llms. *arXiv preprint arXiv:2502.19981*, 2025.

Adib Bazgir, Yuwen Zhang, et al. Agentichypothesis: A survey on hypothesis generation using llm systems. *Towards Agentic AI for Science: Hypothesis Generation, Comprehension, Quantification, and Validation*, 2025a.

Adib Bazgir, Yuwen Zhang, et al. Matagent: A human-in-the-loop multi-agent llm framework for accelerating the material science discovery cycle. In *AI for Accelerated Materials Design-ICLR 2025*, 2025b.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Kourosh Darvish, Marta Skreta, Yuchi Zhao, Naruki Yoshikawa, Sagnik Som, Miroslav Bogdanovic, Yang Cao, Han Hao, Haoping Xu, Alán Aspuru-Guzik, et al. Organa: A robotic assistant for automated chemistry experimentation and characterization. *Matter*, 8(2), 2025.

Joshua Engels, Eric J Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are linear. *arXiv preprint arXiv:2405.14860*, 2024.

Andrew Gambardella, Yusuke Iwasawa, and Yutaka Matsuo. Language models do hard arithmetic tasks easily and hardly do easy arithmetic tasks. *arXiv preprint arXiv:2406.02356*, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Subhash Kantamneni and Max Tegmark. Language models use trigonometry to do addition. *arXiv preprint arXiv:2502.00873*, 2025.

Shrinidhi Kumbhar, Venkatesh Mishra, Kevin Coutinho, Divij Handa, Ashif Iquebal, and Chitta Baral. Hypothesis generation for materials discovery and design using goal-driven and constraint-guided llm agents. *arXiv preprint arXiv:2501.13299*, 2025.

Ge Lei, Ronan Docherty, and Samuel J Cooper. Materials science in the era of large language models: a perspective. *Digital Discovery*, 3(7):1257–1272, 2024.

Quanliang Liu, Maciej P Polak, So Yeon Kim, MD Al Amin Shuvo, Hrishikesh Shridhar Deodhar, Jeongsoo Han, Dane Morgan, and Hyunseok Oh. Beyond designer's knowledge: Generating materials design hypotheses via a large language model. *Acta Materialia*, page 121307, 2025a.

Siyu Liu, Tongqi Wen, Beilin Ye, Zhuoyuan Li, Han Liu, Yang Ren, and David J Srolovitz. Large language models for material property predictions: elastic constant tensor prediction and materials design. *Digital Discovery*, 4(6):1625–1638, 2025b.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Neel Nanda, Senthooran Rajamanoharan, János Kramár, and Rohin Shah. Fact finding: Attempting to reverse-engineer factual recall on the neuron level. `https://www.alignmentforum.org/posts/iGuwZTHWb6DFY3sKB/fact-finding-attempting-to-reverse-engineer-factual-recall`, December 2023. Alignment Forum post, accessed April 29, 2025.

nostalgebraist. Interpreting gpt: The logit lens. `https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens`, August 2020. URL `https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens`. AI Alignment Forum / LessWrong blog post.

Emmanuel A Olowe and Danial Chitnis. Labiium: Ai-enhanced zero-configuration measurement automation system. In *2025 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, pages 1–6. IEEE, 2025.

Jing Qian, Hong Wang, Zekun Li, Shiyang Li, and Xifeng Yan. Limitations of language models in arithmetic and symbolic induction. *arXiv preprint arXiv:2208.05051*, 2022.

Andre Niyongabo Rubungo, Kangming Li, Jason Hattrick-Simpers, and Adji Bousso Dieng. Llm4mat-bench: benchmarking large language models for materials property prediction. *Machine Learning: Science and Technology*, 6(2):020501, 2025.

Malcolm Sim, Mohammad Ghazi Vakili, Felix Strieth-Kalthoff, Han Hao, Riley J Hickman, Santiago Miret, Sergio Pablo-García, and Alán Aspuru-Guzik. Chemos 2.0: An orchestration architecture for chemical self-driving laboratories. *Matter*, 7(9):2959–2977, 2024.

Mark Steyvers, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas W. Mayer, and Padhraic Smyth. What large language models know and what people think they know. *Nature Machine Intelligence*, 7(2):221–231, 2025. doi: 10.1038/s42256-024-00976-7.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Oskar Wysocki, Magdalena Wysocka, Danilo Carvalho, Alex Teodor Bogatu, Danilo Miranda Gusicuma, Maxime Delmas, Harriet Unsworth, and Andre Freitas. An llm-based knowledge synthesis and scientific reasoning framework for biomedical discovery. *arXiv preprint arXiv:2406.18626*, 2024.

Appendix

A Last token residual stream collection

For each layer l, we collect last-token residual streams $\mathbf{h}_{i,j,k}^{(l)}$ from prompts $p_{i,j,k}$ across all elements and templates:

$$\mathbf{h}_{i,j,k}^{(l)} = f^{(l)}\left(p_{i,j,k}\right) \in \mathbb{R}^{T \times d},$$

where $f^{(l)}$ denotes the layer-l transformation, T is the token length, and d is the hidden dimension. The initial residual stream $\mathbf{h}_{i,j,k}^{(0)}$ is obtained by embedding the prompt through an embedding layer E_0, followed by processing through L Transformer layers. Each layer applies multi-head attention and a feedforward network with residual connections and layer normalization:

$$\begin{aligned}
\mathbf{h}_{i,j,k}^{(l)\prime} &= \mathbf{h}_{i,j,k}^{(l-1)} + \text{Attention}\big(\text{Norm}(\mathbf{h}_{i,j,k}^{(l-1)})\big) \\
\mathbf{h}_{i,j,k}^{(l)} &= \mathbf{h}_{i,j,k}^{(l)\prime} + \text{FFN}\big(\text{Norm}(\mathbf{h}_{i,j,k}^{(l)\prime})\big)
\end{aligned}$$

Here, $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ represent the query, key, and value matrices used in multi-head attention to compute token-to-token interactions. Finally, $\mathbf{h}_{i,j,k}^{(L)}$ is mapped to the vocabulary space using the vocabulary head $\mathbf{W}_{\text{vocab}}$ to produce logits:

$$\text{logits}_{i,j,k} = \mathbf{h}_{i,j,k}^{(L)} \mathbf{W}_{\text{vocab}}$$

By analyzing last-token residual streams $\mathbf{h}_{i,j,k}^{(l)}$ across layers, we investigate how attributes are represented in the model's hidden states.

B Intervention outcomes in geometric recall

B.1 Layer-wise performance evaluation

Fig. B.1 illustrates the prediction error across layers when the residual stream of the last token across layers is replaced with the predicted residual stream derived from the geometric space $f(r, g, p) = (r\cos\theta, r\sin\theta, r)$. In the early layers, errors gradually decrease because the model has not yet captured semantic information, and the geometric space is still being constructed. The continuous decline in error reflects the model's growing ability to capture semantic information and progressively build a coherent geometric representation. By layer 20, the error stabilizes, indicating that these layers effectively encode the periodic and geometric relationships between atomic properties such as atomic number, group, and period.

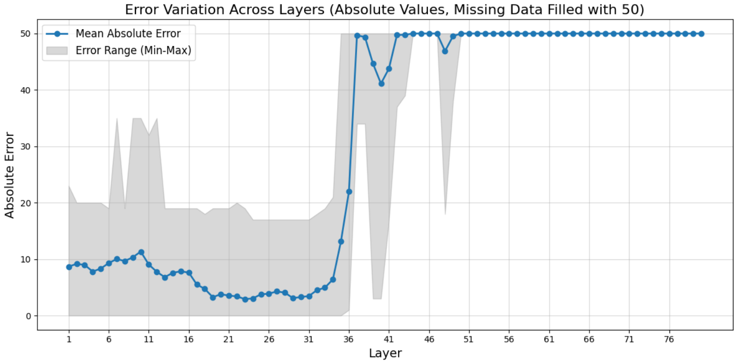


Figure B.1: Variation of Absolute Prediction Errors Across Layers with Intervention. The plot shows the mean absolute error (MAE) for each layer, along with the minimum and maximum error range represented by the shaded region. Missing data points were replaced with a value of 50 before computing the absolute errors.

However, beyond layer 30, the error increases sharply as the model begins outputting non-numeric tokens (replaced with an error value of 50 in the plot, corresponding to the maximum possible error given the 50 elements and atomic numbers). This can be attributed to two factors. First, if the numeric token is not the first output, generating the correct answer requires residual streams across all token positions. But only the last residual stream was replaced. Intervening too late disrupts the established flow of residual streams at other positions, which have already determined the output content. Second, it is also likely due to the model shifting its focus from geometric relationships to higher-level abstractions or context-dependent reasoning in the later layers. Therefore, for the intervention experiments on geometric relationships, we selected layer 20 as it balances effective encoding of geometric relationships and minimizes disruption to the model's output process.

B.2 Detailed evaluation of geometric spaces

The primary evaluation criterion used in the main text is the absolute error threshold (≤ 2), as discussed in detail in Sec.3. This metric was chosen because it better captures the accuracy of residual stream interventions. However, other metrics, such as R^2, Pearson correlation, and qualitative mapping fidelity, also provide valuable insights. These complementary results are summarized in Table B.1.

#	Space	Description	$\mathbf{R^2}$	Pearson Correlation	Percentage of Abs. err ≤ 2	Mapping Fidelity
1	r	Linear structure along atomic number.	0.8863	0.9591	38.00%	Moderate
2	(r, g, p)	3D cartesian grid.	0.8060	0.9191	48.00%	Moderate
3	$(r\cos\theta, r\sin\theta, r)$	3D radial spiral structure.	0.8162	0.9035	72.00%	High
4	$(\cos\theta, \sin\theta, r)$	3D spiral structure.	0.7596	0.8813	70.00%	High
5	$(\cos\theta, \sin\theta, p)$	3D periodic wave-like structure.	0.5106	0.7174	60.00%	Moderate
6	$(r\cos\theta, r\sin\theta, p)$	3D periodic lattice with radial dependencies.	0.6719	0.8240	62.00%	Moderate
7	$(r\cos\theta, r\sin\theta)$	2D radial structure.	-0.1391	0.1481	40.00%	Low
8	r_{random}	Random linear structure.	0.0075	0.1503	10.00%	Low
9	$(\cos(\theta_{\text{random}}), \sin(\theta_{\text{random}}), r)$	Randomized spiral.	0.6358	0.8465	20.00%	Low
10	$(r\cos\theta, r\sin\theta, r)$	Element unrelated prompts	-0.4910	0.7215	48.00%	Low

Table B.1: Performance of different low-dimensional spaces for residual stream intervention. Each space represents a unique pattern, with results assessed using R^2, Pearson correlation, and percentage of predictions within absolute error ≤ 2.

In the main paper, we demonstrate two geometric space intervention results; however, other shapes can also be extracted. Fig. B.2 shows the extracted linear structure from interventions. While the alignment of points along a straight path indicates the presence of a linear structure, the overlapping points suggest its limitations in distinguishing atomic number. Compared to more expressive shapes like spirals, linear structures may struggle to effectively capture periodic or distinct features.

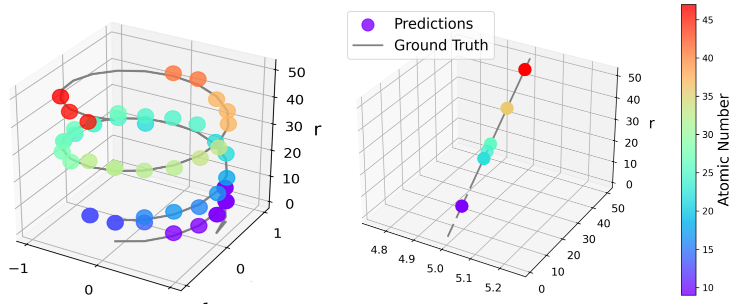


Figure B.2: Helix and Linear Structure in the Geometric Space from Intervention Experiments. The figure shows predictions (colored points) and their alignment with the ground truth (gray line).

C Attention map detailed results

To investigate how the model prioritizes different parts of the input text, we conducted a preliminary analysis using the 32-layer Meta-Llama-3-8B model. We adopted the attribute A_j, Period and Group, and iterated over X_i, consisting of 50 elements, using the prompt template: 'In the periodic table of elements, the A_j of X_i is.' These prompts were input into the language model, and we analyzed the average attention across all attention heads in each transformer layer from the token 'is' to all other tokens. The averaged results across different prompts are presented in Fig.C.1.

The results indicate that in the intermediate layers, where entropy is relatively high, there is a noticeable concentration of attention from the token 'is' to attribute and element tokens. This suggests that these intermediate layers focus more on tokens within the sequence that have a significant impact on the output. In contrast, the later layers, which exhibit lower entropy (with the exception of the final layer), show a more evenly distributed attention pattern. This pattern implies that the model transitions from focusing on specific token relationships to integrating broader context, thereby finalizing its interpretation for a cohesive output.

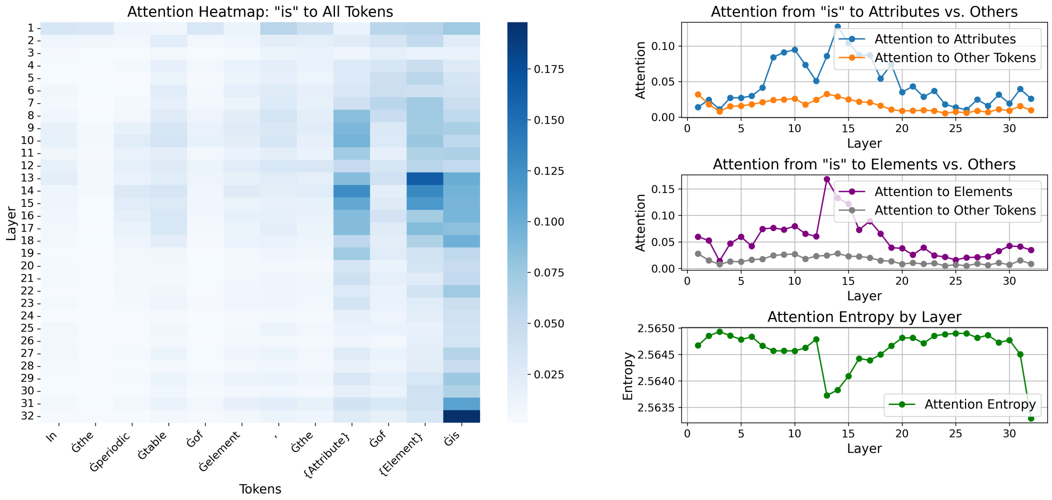


Figure C.1: Average attention distribution analysis of the 32-layer Meta-Llama-3-8B model across transformer layers, based on prompts, 'In the periodic table of elements, the A_j of X_i is,' where A_j is an attribute (period or group) and X_i is an element. The heatmap (left) shows average attention from 'is' to all tokens, while line plots (right) depict attention to target tokens (e.g., element and attribute), average attention to other tokens, and attention entropy. Intermediate layers focus on significant token relationships with higher entropy, while later layers (excluding the final layer) show evenly distributed attention and lower entropy, reflecting a shift to broader context integration.

D Logit lens and tuned lens

We input the prompt 'The atomic number of Mg is ' and analyzed the token probabilities at each layer using logit-len. By normalizing the final token's hidden state with LayerNorm and applying the vocabulary head followed by softmax, we obtained the top-ranked tokens directly output by each layer. In each layer, we extracted the probability of the target token, t_{target} — the output token from the last layer, and checked if it ranked within the top 50 most probable tokens for that layer. The results are shown in Fig. D.1.

In the early layers, the probability of the target token has not shown an upward trend, indicating these layers neither strongly predict the target tokens nor significantly refine their probabilities. In contrast, probabilities gradually increase in the later layers, highlighting their role in refining and finalizing predictions. Although crucially, there don't appear to be any hard boundaries between these distinct activities and the model smoothly transition from one to the next. The markers, concentrated in later layers, suggest that while intermediate layers store factual knowledge, they are not yet attempting to articulate it in language form.

Notably, the distribution of 'Top 50' markers varies by token type. Tokens with lower contextual complexity, such as spaces, 'and,' or 'since,' have their markers in earlier layers. In contrast, knowledge-based tokens, like '12,' require deeper processing and appear in much later layers. This suggests that while intermediate layers encode factual concepts, they are likely focused on tasks other than linguistic articulation, which primarily develops in the later layers.

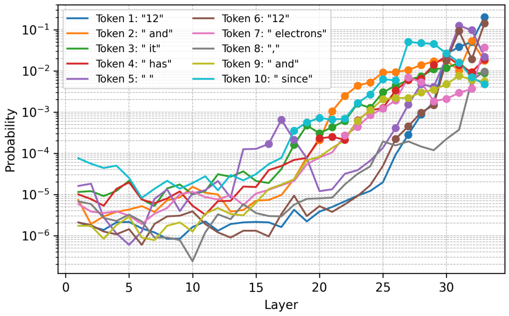


Figure D.1: Probability of the target token t_{target} across layers in Meta-Llama-3-8B for the prompt "The atomic number of Mg is". Following the logit lens framework, each line shows t_{target}'s probability derived from intermediate layer logits. Probabilities are computed by iteratively re-running the model with the next token added. Markers indicate layers where t_{target} ranks in the top 50 most probable tokens.

We also performed additional experiments using tuned lens in Fig D.2. The results consistently show a similar trend: tokens such as "of" and "is" achieve high prediction accuracy at intermediate or earlier layers, while more factual or critical token (e.g., blue line "5" only becomes the top prediction at layer 29)—require deeper processing.

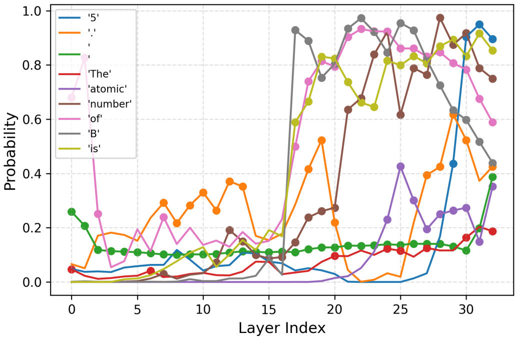


Figure D.2: Probability of the target token t_{target} across layers in Meta-Llama-3-8B for the prompt "The atomic number of Mg is". Each line shows t_{target}'s probability computed using a tuned lens—a learned linear probe trained to decode intermediate hidden states. Probabilities are obtained by iteratively re-running the model with the next token appended. Markers indicate layers where t_{target} ranks among the top prediction.

E Attribute representations overlap in intermediate layers

E.1 Probing weights analysis

As outlined in Sec. 4.1, we trained a linear model for each attribute A_j at each layer l, yielding a weight vector $\mathbf{w}_j^{(l)}$ that represents how attribute A_j is stored in the residual stream space of layer l. To analyze attribute relationships across layers, we computed the cosine similarity between weight vectors of different attributes using continuation-style residual stream sets to minimize language pattern influence.

Fig. E.1 illustrates the cosine similarity across 80 layers of Meta-Llama-3.1-70B. Notably, in high-dimensional spaces, random vector pairs typically approach orthogonality due to the 'blessing of dimensionality'. To illustrate this, we randomly sampled vector pairs in an 8129-dimensional space (the residual stream vector size of Meta-Llama-3.1-70B) and calculated their cosine similarity, with

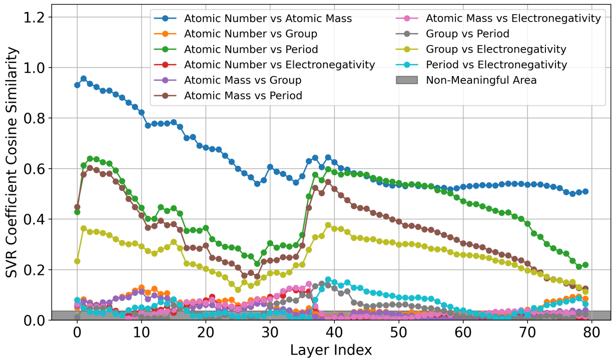


Figure E.1: Cosine similarity between weight vectors of linear probes for attribute pairs across layers in Meta-Llama-3.1-70B. The shaded area (99.9% CI) represents unrelated boundaries.

the 99.9% confidence interval (CI) shown in gray. Cosine similarity outside this interval indicates meaningful relationships between attributes. See Appendix E.2 for more details.

In the early layers, high similarity reflects token-level processing rather than semantic understanding. As layers deepen, similarity decreases as the model begins capturing semantics. In the intermediate layers, similarity rises, indicating shared representation of correlated attributes. Finally, in the later layers, similarity drops again as the model separates features for refined decision-making.

E.2 Blessing of dimensionality

When the dimensionality is very high, the most of random vector pairs approach orthogonality. We illustrate this by sampling pairs of vectors in an 8129-dimensional space (corresponding to the residual stream vector dimension of Meta-Llama-3.1-70B) and computing their cosine similarities. The 99.9% confidence interval (CI) provides an estimate of the expected cosine similarity range at each dimensionality:

$$\mathrm{CI}_{99.9\%} = \left(\mu - z\frac{\sigma}{\sqrt{n}}, \mu + z\frac{\sigma}{\sqrt{n}}\right)$$

where μ is the sample mean, σ is the sample standard deviation, n is the number of sampled pairs, and $z \approx 3.29$ for a 99.9% confidence level.

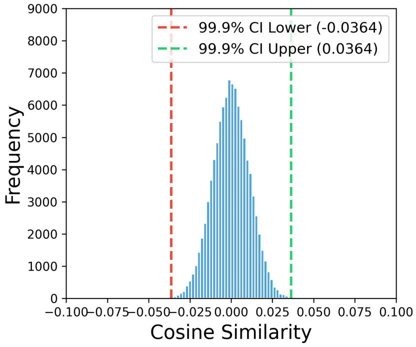


Figure E.2: Cosine similarity distribution of random vector pairs in an 8129-dimensional space, with a 99.9% confidence interval $(-0.0364, 0.0364)$ shown by the dashed lines.

The specific distribution is shown in Fig. E.1, where the confidence interval is extremely narrow (± 0.036), indicating that random vector pairs exhibit highly consistent cosine similarities. This suggests that the learned weights of the linear probe across different feature pairs are effectively uncorrelated, exhibiting only random alignment. In Fig. 3, the shaded region represents the 99.9%

confidence interval for the cosine similarities of high-dimensional random vectors, further supporting this observation.

F Linear probing detailed results

In Sec.4.1, we applied linear probing to train linear-kernel Support Vector Regression (SVR) models for each layer l and attribute A_j, using the residual stream dataset $H_j(l)$. The ground truth values $y_{i,j}$ correspond to the attribute values of each element X_i.

F.1 Why R^2 cannot reach 1

Even with a perfectly trained model and a sufficiently large dataset, achieving an R^2 of 1 in linear probing is impossible. The model's output token '1' indicates that the residual stream at the last token position in the final layer leads to the highest logit for token 1's ID after the final linear transformation. However, do token residual stream in the final layer of numbers exhibit a perfect linear relationship with their real numerical values? Token embeddings are learned representations that capture semantic relationships between tokens, but they are not guaranteed to align linearly with numerical values.

To further investigate this, we fit a linear model to map approximated numerical token representations in the last layer to their actual values. Specifically, we extract token IDs for numbers 1–50 from the tokenizer, multiply them by the pseudoinverse of the vocabulary projection matrix $\mathbf{W}_{\text{vocab}}^{+}$, and obtain their corresponding vector representations in the hidden space of the last layer:

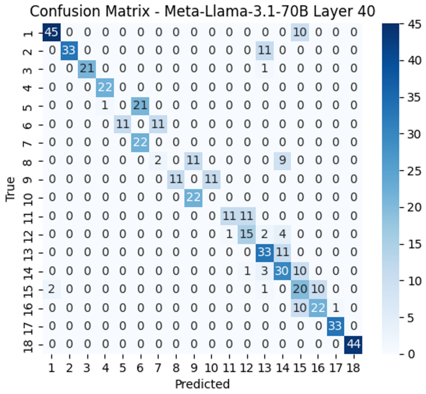
$$\mathbf{h}_i = \mathbf{W}_{\text{vocab}}^{+} \cdot \mathbf{t}_i$$

where $\mathbf{t}_i$ is the encoded token ID for the number i, and $\mathbf{h}_i$ represents its corresponding hidden space representation.

We then fit a linear regression model to map these representations to their true numerical values. The linear correlation turned out to be quite strong, with an R^2 of 0.98. However, this is not 1—possibly because the embedding space is not perfectly linearly aligned with numerical values, or because it is influenced by semantic noise, or simply due to limitations in the fitting method.

For LLMs, even if the logits were identical to the embeddings $\mathbf{h}_i$ (which is theoretically impossible—at best, they can only approximate them), the R^2 would still be limited to 0.98. Therefore, it is unsurprising that linear probing does not achieve an R^2 of 1.

F.2 Confusion matrix



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	45	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0
2	0	33	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0
3	0	0	21	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
4	0	0	0	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	1	0	21	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	11	0	11	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	22	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	2	0	11	0	0	0	0	9	0	0	0	0
9	0	0	0	0	0	0	0	11	0	11	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	22	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	11	11	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	1	15	2	4	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	33	11	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	1	3	30	10	0	0	0
15	2	0	0	0	0	0	0	0	0	0	0	0	1	0	20	10	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	22	1	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	33	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	44

Figure F.1: Confusion Matrix on categorical linear probing of attribute 'Group' on the middle layer.

F.3 Question prompts

We applied the Mann–Kendall test to ΔR^2 in the upper 50% of layers for each model–attribute pair (15 in total). Benjamini–Hochberg adjustment (FDR = 0.05) retained significance in 12 pairs, with τ ranging from 0.36 to 0.94 (median 0.79; Table F.3).

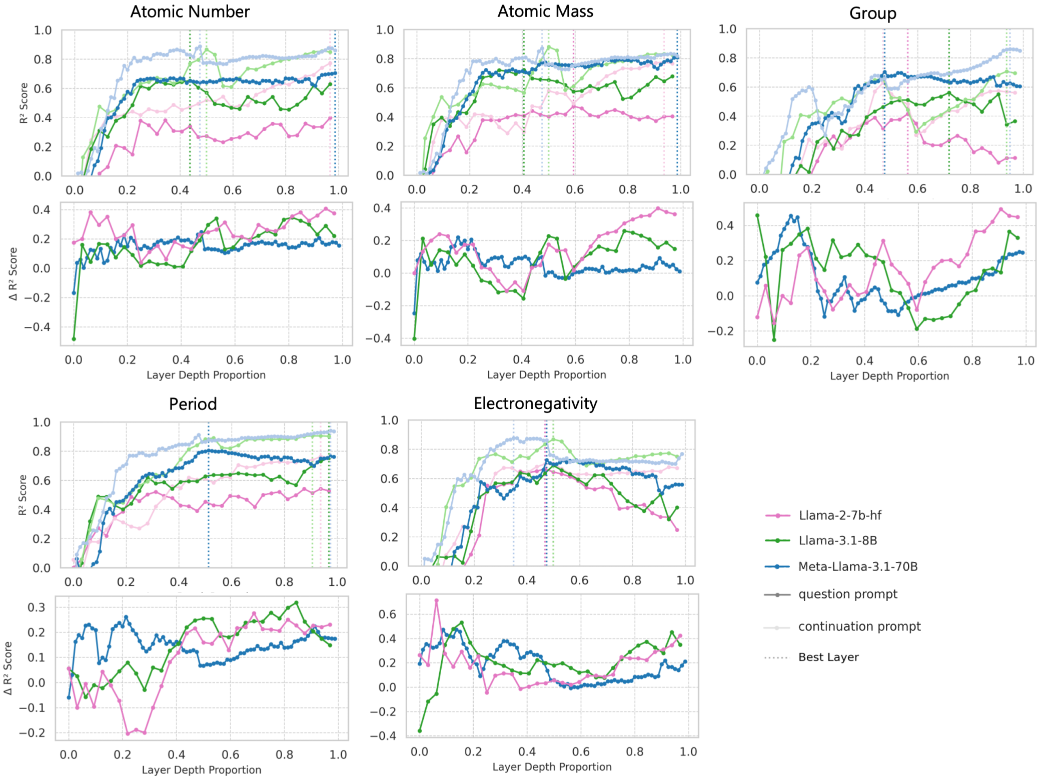


Figure F.2: R² scores for linear probes trained on target properties and evaluated on representations from question and continuation prompts.

Table F.1: Mann–Kendall trend test on ΔR^2 (difference between continuation and question prompt R²) within depth layer depths 0.5–1.0. τ: Kendall's tau (trend strength). (*) indicates significance after Benjamini–Hochberg FDR correction ($\alpha = 0.05$).

Model	Attribute	τ	p
Meta–Llama–3.1–70B	Group	0.95	$<$ 0.001*
	Period	0.84	$<$ 0.001*
	Electronegativity	0.73	$<$ 0.001*
	Atomic Mass	0.53	$<$ 0.001*
	Atomic Number	0.41	$<$ 0.001*
Llama–2–7b–hf	Electronegativity	0.80	$<$ 0.001*
	Atomic Mass	0.80	$<$ 0.001*
	Group	0.75	$<$ 0.001*
	Atomic Number	0.55	0.003*
	Period	0.40	0.034*
Llama–3.1–8B	Group	0.57	0.003*
	Electronegativity	0.52	0.006*
	Atomic Number	0.20	0.300
	Atomic Mass	0.20	0.300
	Period	-0.05	0.822

F.4 ‘Non-matching’ and ‘no mention’ prompts

F.4.1 Non-matching attribute pairs

To contrast explicit and implicit attribute cues, we selected a set of *non-matching* attribute pairs with minimal direct statistical dependency. Each pair was chosen to lack strong linear or monotonic relationships, and we verified on our 50-element subset that all selected pairs satisfy $|\text{Pearson}\, r| < 0.30$, $|\text{Spearman}\, \rho| < 0.30$, and $R^2 < 0.15$ (Table F.2). These thresholds ensure that no pair can be trivially recovered through simple linear or monotonic mappings.

Table F.2: Linear and monotonic correlations for non-matching attribute pairs (50-element subset).

Pair	|Pearson r|	|Spearman ρ|	Linear R²
Group–Atomic Number	0.044	0.070	0.002
Group–Period	0.255	0.300	0.065
Group–Mass	0.037	0.066	0.001
Electronegativity–Atomic Number	0.154	0.038	0.024
Electronegativity–Mass	0.147	0.039	0.022

F.4.2 Detailed results

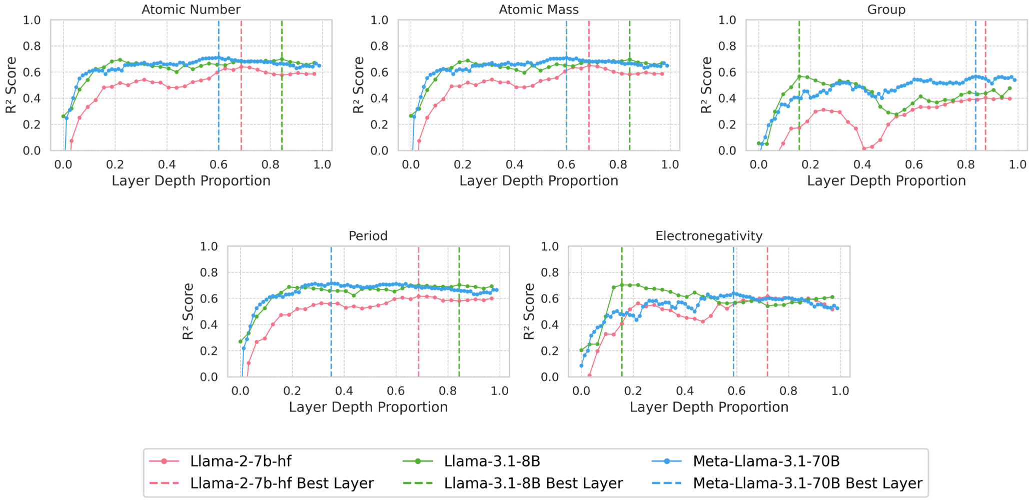


Figure F.3: R² score trends for ‘no mention’ cases. Regression linear probing on the element token residual stream with 5-fold cross-validation was performed on residual streams, and R2 scores on the test set are shown for each attribute.

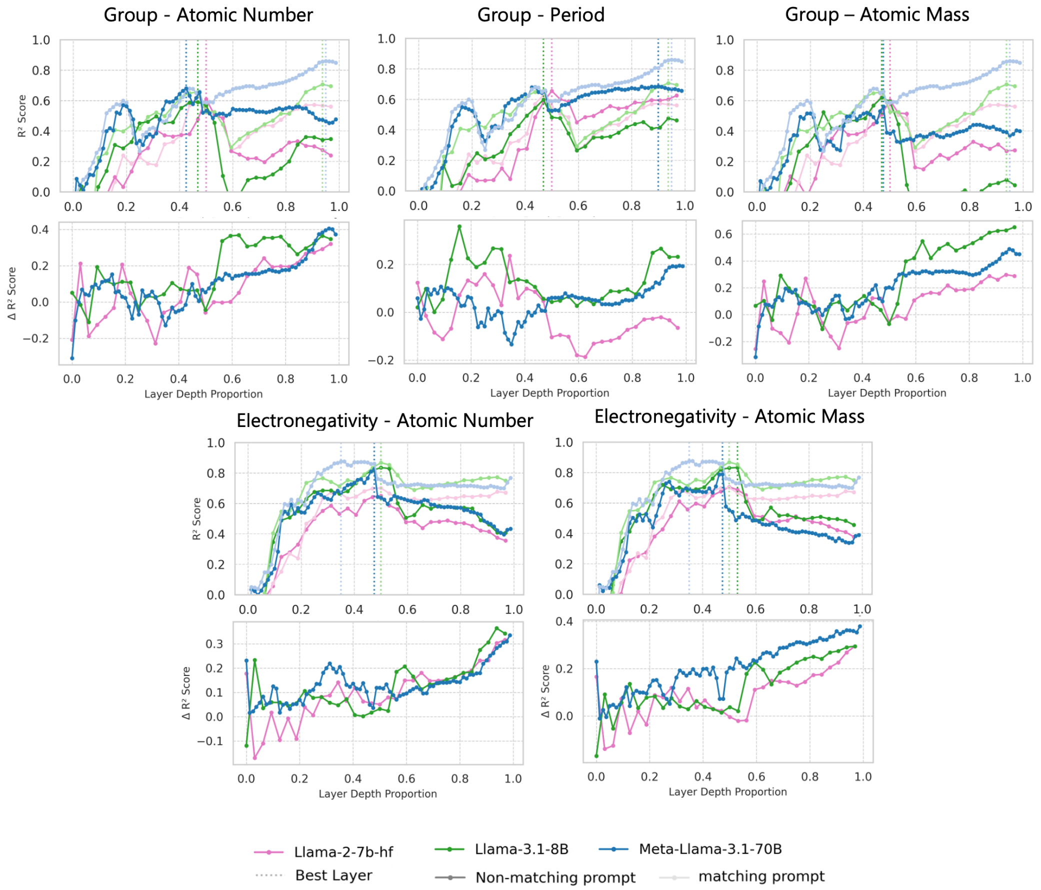


Figure F.4: R² and ΔR^2 scores for linear probes trained on target properties and evaluated on representations from matching and non-matching prompts. ΔR is defined as R²(matching prompt) minus R²(non-matching prompt).

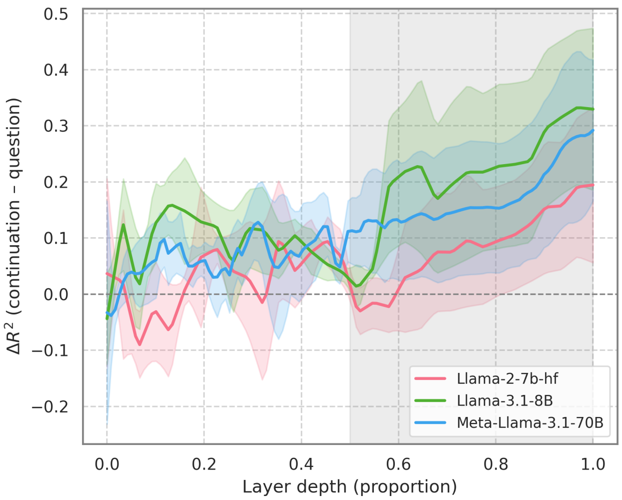


Figure F.5: Average ΔR^2 across five attributes, with with 95% confidence interval shaded. $\Delta R^2 = R^2_{\text{match}} - R^2_{\text{non-match}}$.

Table F.3: Mann–Kendall trend test on ΔR^2 (difference between matching and non-matching prompt R²) within layer depths 0.6–1.0. τ: Kendall's tau (trend strength). "*" indicates significance after Benjamini–Hochberg FDR correction ($\alpha = 0.05$).

Model	Attribute Pair	τ	p
Meta–Llama–3.1–70B	Electronegativity – Atomic Number	0.923	< 0.001*
	Group – Atomic Number	0.911	< 0.001*
	Electronegativity – Atomic Mass	0.895	< 0.001*
	Group – Period	0.625	< 0.001*
	Group – Atomic Mass	0.452	< 0.001*
Llama–2–7b–hf	Group – Atomic Number	0.872	< 0.001*
	Group – Atomic Mass	0.846	< 0.001*
	Electronegativity – Atomic Mass	0.795	< 0.001*
	Group – Period	0.769	< 0.001*
	Electronegativity – Atomic Number	0.769	< 0.001*
Llama–3.1–8B	Electronegativity – Atomic Mass	0.821	< 0.001*
	Group – Period	0.769	< 0.001*
	Group – Atomic Mass	0.744	< 0.001*
	Electronegativity – Atomic Number	0.641	0.003*
	Group – Atomic Number	-0.154	0.502

F.5 Detailed results of the best layer

In the main text Sec. 4.1, we used SVR for linear regression probing. Figures F.6, F.7, F.8, F.9, and F.10 present the detailed R^2 performance of the best layer for each attribute—atomic number, atomic mass, electronegativity, period, and group—across three different models.

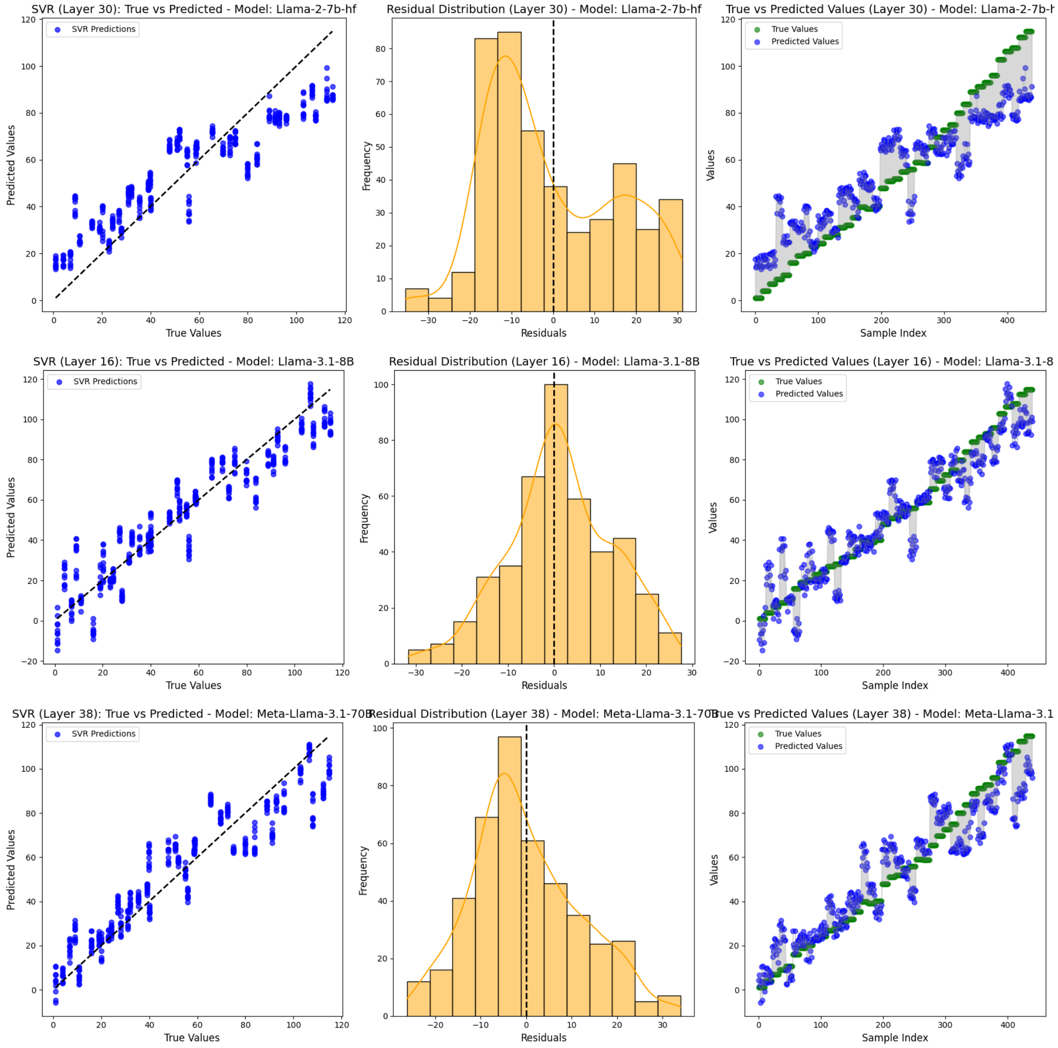


Figure F.6: Evaluation of SVR performance for Layer best_layer on the atomic mass. The left plot shows true vs. predicted values with alignment to the diagonal indicating accuracy. The center plot displays residuals, highlighting error distribution centered around zero. The right plot visualizes true and predicted values across samples, with shaded areas representing error magnitudes.

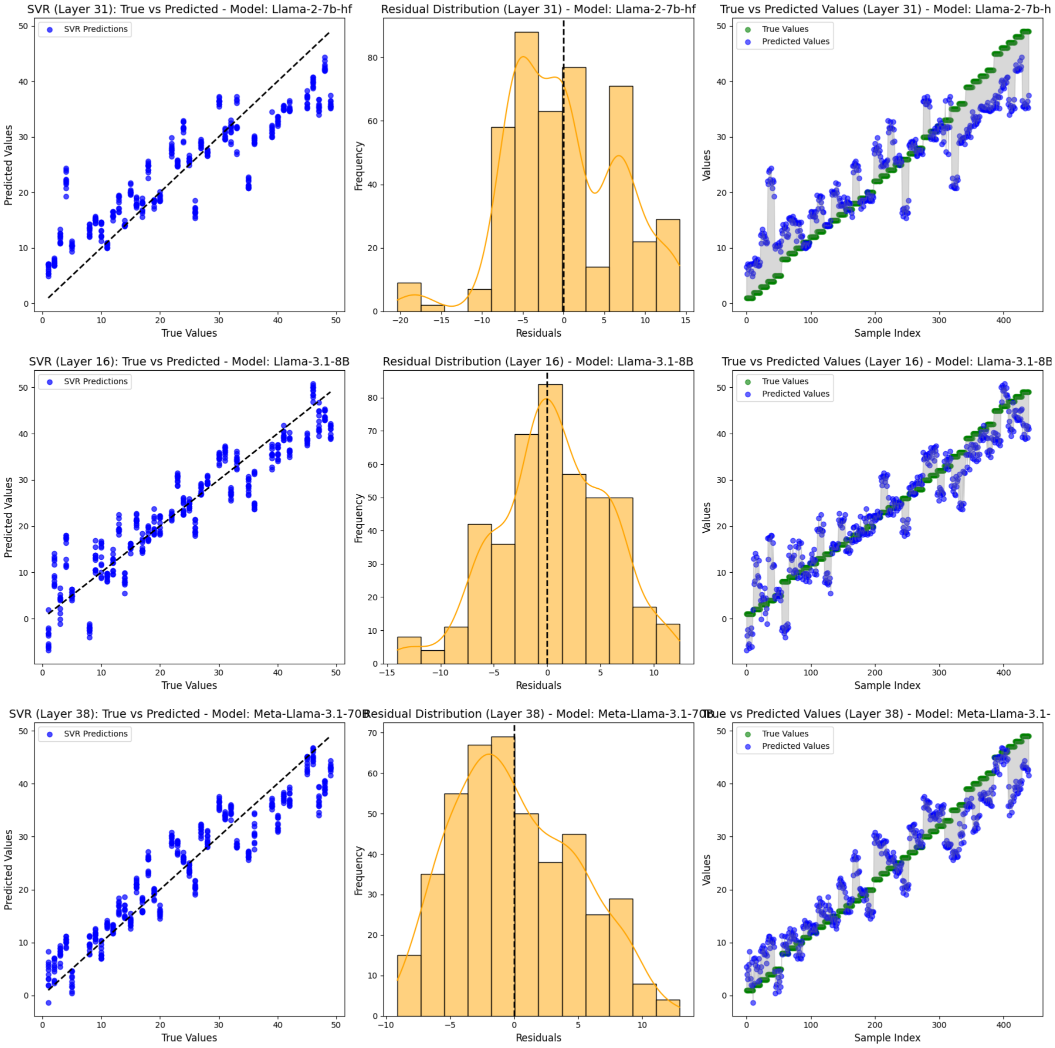


Figure F.7: Evaluation of SVR performance for Layer best_layer on the atomic number. The left plot shows true vs. predicted values with alignment to the diagonal indicating accuracy. The center plot displays residuals, highlighting error distribution centered around zero. The right plot visualizes true and predicted values across samples, with shaded areas representing error magnitudes.

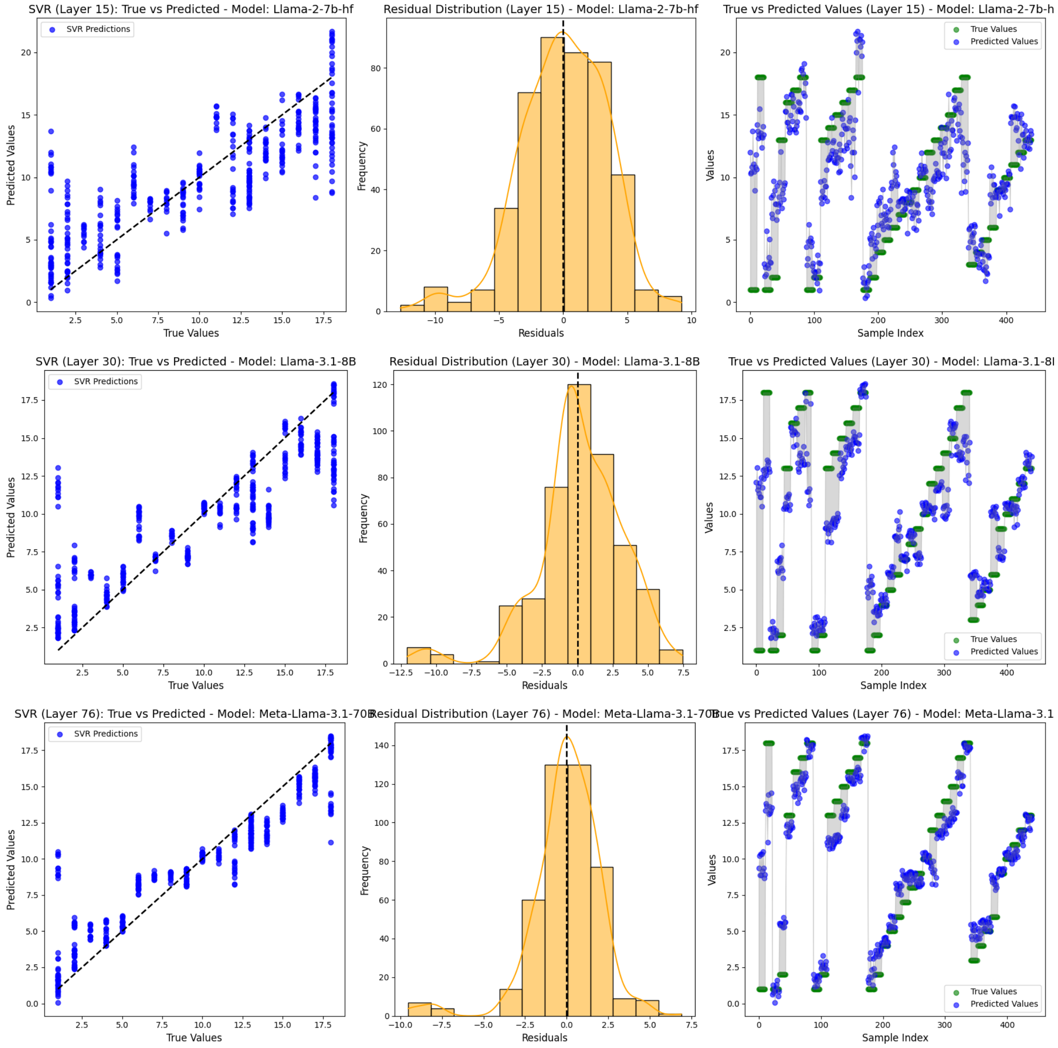
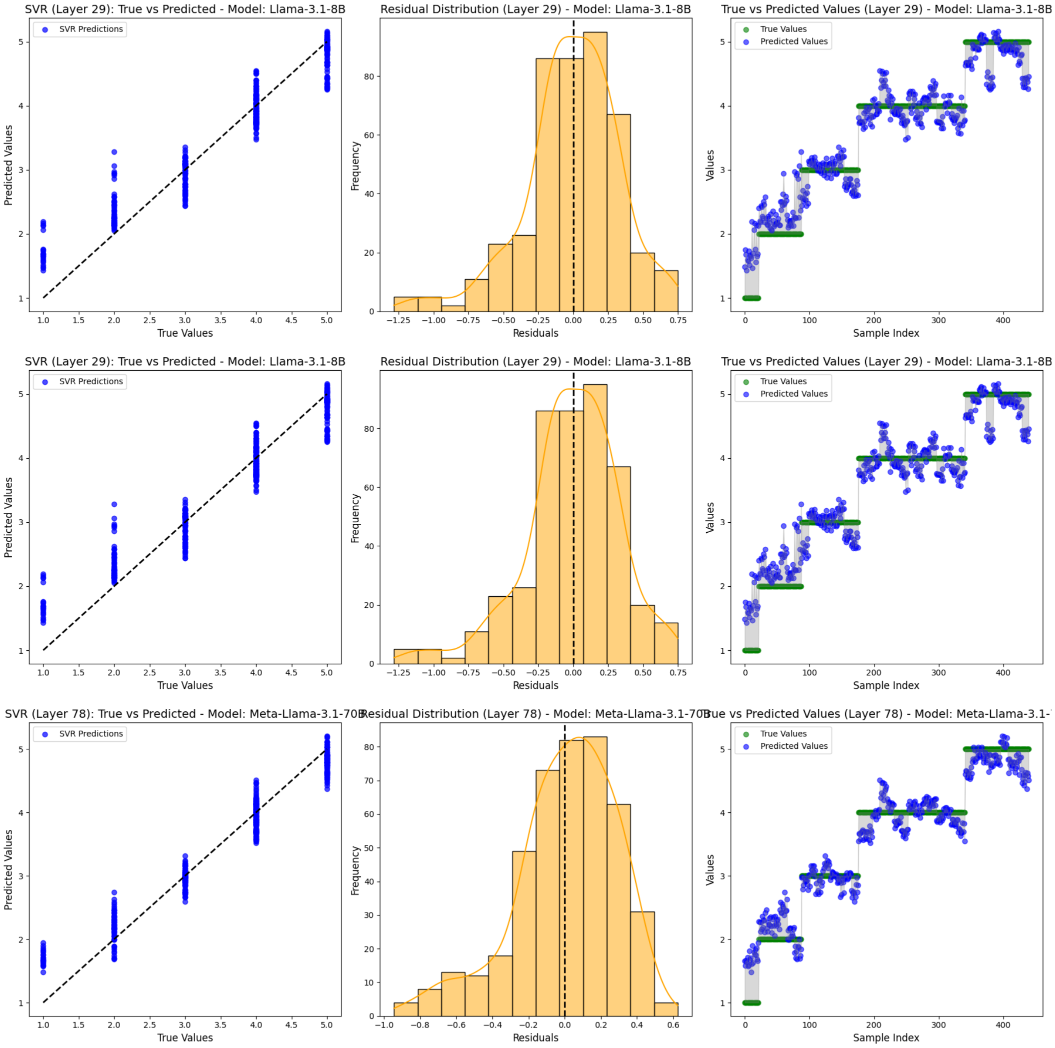


Figure F.8: Evaluation of SVR performance for Layer best_layer on the group. The left plot shows true vs. predicted values with alignment to the diagonal indicating accuracy. The center plot displays residuals, highlighting error distribution centered around zero. The right plot visualizes true and predicted values across samples, with shaded areas representing error magnitudes.


SVR (Layer 29): True vs Predicted - Model: Llama-3.1-8B

Residual Distribution (Layer 29) - Model: Llama-3.1-8B

True vs Predicted Values (Layer 29) - Model: Llama-3.1-8B

SVR (Layer 29): True vs Predicted - Model: Llama-3.1-8B

Residual Distribution (Layer 29) - Model: Llama-3.1-8B

True vs Predicted Values (Layer 29) - Model: Llama-3.1-8B

SVR (Layer 78): True vs Predicted - Model: Meta-Llama-3.1-70B

Residual Distribution (Layer 78) - Model: Meta-Llama-3.1-70B

True vs Predicted Values (Layer 78) - Model: Meta-Llama-3.1-

Figure F.9: Evaluation of SVR performance for Layer best_layer on the period. The left plot shows true vs. predicted values with alignment to the diagonal indicating accuracy. The center plot displays residuals, highlighting error distribution centered around zero. The right plot visualizes true and predicted values across samples, with shaded areas representing error magnitudes.

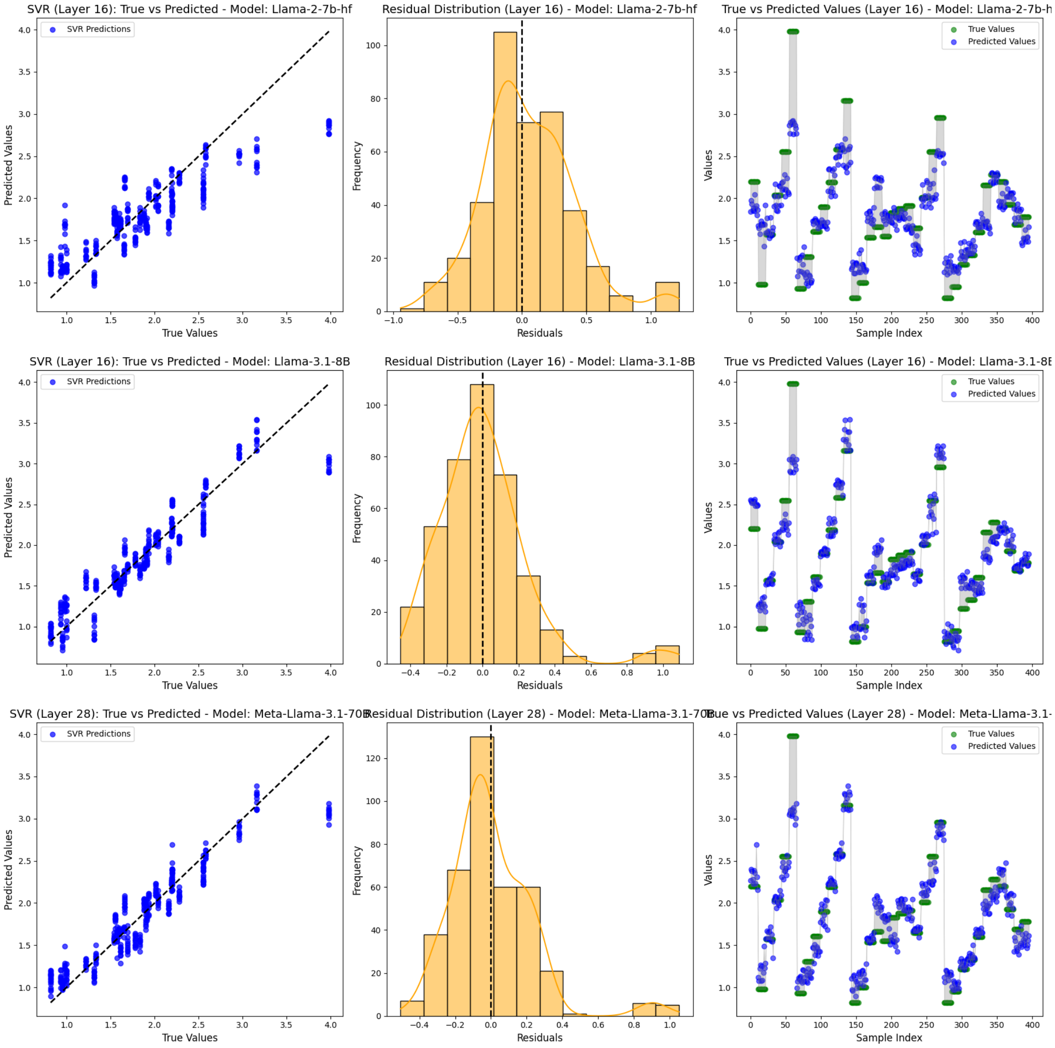


Figure F.10: Evaluation of SVR performance for Layer best_layer on the electronegativity. The left plot shows true vs. predicted values with alignment to the diagonal indicating accuracy. The center plot displays residuals, highlighting error distribution centered around zero. The right plot visualizes true and predicted values across samples, with shaded areas representing error magnitudes.